Testing AI:

How do we improve LLM safety, helpfulness, and truthfulness?

Authors: Peter Kant and Kiersten Todt

Enabled Intelligence

Overview

Artificial intelligence holds incredible promise to simplify our lives by taking meeting notes, creating first drafts of correspondence, identifying objects in the real world to power driverless cars, and giving our military the ability to better understand situations around the globe as they unfold.

While there is great excitement about this promise, there are also legitimate concerns regarding the safety, helpfulness, and truthfulness of these platforms. Before the general public can rely on AI as a useful tool, AI-powered large language models (LLMs) require more testing and fine tuning. Testing LLMs is much more complicated than testing other technological functionality. Such tests require the evaluation of tone, context, and emotional impact. They also require an understanding of how LLMs work and how they make mistakes. Both elements render human testers essential – a fact that many AI companies are reluctant to admit.

LLMs work similar to the way your smartphone does when it helps you complete your text message: It makes a prediction about what words should come next. The answers provided by the LLM are only predictions, and the prediction is made using whatever dataset it was trained on – often information found on the internet or a large, purchased dataset that includes relevant information. The LLM then generates a response that has the highest probability of accuracy based on the information it was trained on. However, if the dataset that was used to train the LLM was not diverse or large enough, the margin of error can be significant.

Because of their predictive power, LLMs are very good at generating answers that sound completely accurate but are factually incorrect or even completely made up. This phenomenon is called hallucination. And while this result can be amusing in some situations, it can be lethal in other contexts. Imagine, for example, a hallucination scenario playing out for a counter-terrorism analyst at the Department of Homeland Security or a cancer drug researcher.

As we begin to deploy AI in service of the warfighter and in other high consequence applications, we must invest in rigorous human testing of our LLMs. When speed and precision are paramount, proper training, testing, and evaluation can literally be life or death.

> ***"Much greater investments are needed in [DoD] AI testing and evaluation [T&E] than previous T&E resources; partially because previous T&E was notoriously under-resourced and because AI systems are so complex."*** [1]

Evaluating Safety

We are consistently reminded that those looking to cause harm often go to the Internet first. Nefarious actors use the Internet to find information on how to build weapons, create and spread disinformation, and promote bias. LLMs do not yet "know" how to tell bad actors and violent, racist, misogynistic, and unsafe content from other content. LLMs currently "learn" from this biased, inaccurate content along with the rest of the data collected from open sources during training. As recent horrific examples from Google, Microsoft and other LLM launches show, even the most sophisticated LLMs quickly start providing harmful, hateful, and violent content. In these examples, human intervention is critical to root out these harmful LLM developments. Human testers are crucial in understanding when LLM generated responses or information are potentially harmful or unsafe. For example, is the LLM providing helpful women's health information or sexual assault care access information? Or is it describing sexual violence or using misogynistic language? Is the LLM providing information on historical racial discrimination? Or is it parroting racist content learned from social media? The words and language are not that different, but the context and human understanding is absolutely necessary in determining the intent and safety of the content.

Employing a Comprehensive Evaluation Ontology

Most LLM evaluations use a basic ontology of assessing "Helpfulness" (does the response follow the prompt instructions and provide an appropriate response); "Truthfulness" (does the response contain any factual errors or hallucinations); and "Safety" (does the response contain harmful, racist, hateful, pornographic, or violent language or direction). This ontology is often not enough. Ontologies need to capture edge cases (would a response detailing reproductive health options be considered "sexual" or "pornographic"?; would a response providing a history of the Jim Crow era be classified as "racist?"), and other concepts like tone, style, verbosity, naturalness of language. Ontology must also cover the interactions of the LLM: Is the LLM designed to sound "human-like" or should the LLM present itself as a technical assistant, and respond more like a machine?

1 National Academy of Sciences, "Test and Evaluation Challenges in Artificial Intelligence-Enabled Systems for the Department of the Air Force (2023)"

Ensuring Helpfulness + Truthfulness

While it might seem cost-effective to offshore LLM testing and evaluation, doing so is a short-sighted economic calculus that ends up being more expensive – in financial and social costs – in the long run. Evaluating an LLM requires a strong competency in the language and related culture in which prompts are given. Off-shore gig workers may not have the skills required to accurately assess English language LLM results, lacking an understanding of common turns of phrase, euphemisms, and other nuances that are not readily apparent to non-native English language speakers.

Using diverse evaluators to spot bias is another way to ensure accuracy. Detecting bias is more than identifying racist or misogynistic language. Bias detection requires diversity of understanding, thought, and experience. For example, a gig worker in India may not know the history of financial "redlining" that prevented many people of color from getting mortgages. That tester would not be able to recognize when an LLM inaccurately recommends against investing in real estate in more diverse areas based on past mortgage approval rates.

In Conclusion: The Human Factor in AI Testing and Evaluation

AI holds unique and unparalleled promise, but this promise can only be realized if the time, effort, and investment is invested to ensure these platforms are helpful, truthful, and safe. This challenge is not only about technology. AI requires expert humans in the loop, and human training is an indispensable if often overlooked prerequisite to the success of AI.[2] As we look to improve AI platforms, and ensure they are constructed to create positive societal change, we need to remember that expert and well-trained humans are at the core of this important equation.

2 National Academy of Sciences: "As demonstrated by previous examples of AI projects carried out at scale and both DoD – and industry-wide digital modernization programs, leaders commonly underestimate the investments of … expert human resources … required to establish modern AI data best practices."


Enabled Intelligence